MM-LLMs: Recent Advances in MultiModal Large Language Models

Abstract

In the past year, MultiModal Large Language Models (MM-LLMs) have undergone substantial advancements, augmenting off-the-shelf LLMs to support MM inputs or outputs via cost-effective training strategies. The resulting models not only preserve the inherent reasoning and decision-making capabilities of LLMs but also empower a diverse range of MM tasks. In this paper, we provide a comprehensive survey aimed at facilitating further research of MM-LLMs. Initially, we outline general design formulations for model architecture and training pipeline. Subsequently, we introduce a taxonomy encompassing 126 MM-LLMs, each characterized by its specific formulations. Furthermore, we review the performance of selected MM-LLMs on mainstream benchmarks and summarize key training recipes to enhance the potency of MM-LLMs. Finally, we explore promising directions for MM-LLMs while concurrently maintaining a real-time tracking website for the latest developments in the field. We hope that this survey contributes to the ongoing advancement of the MM-LLMs domain.